

AD-A107 816

HARVARD UNIV CAMBRIDGE MASS DEPT OF STATISTICS
DENSITY ESTIMATION AND PROJECTION PURSUIT METHODS. (U)
SEP 81 P J HUBER

F/6 12/1

N00014-79-C-0512

NL

UNCLASSIFIED PJH-7

[OF]

AD A
107816



END
DATE
FILMED
11-82
DTIC

AD A107816

20

LEVEL

RESEARCH REPORT

PRI - 7

DENSITY ESTIMATION AND PROJECTION PURSUIT METHODS

Peter J. Huber
September 1981

Density estimation and projection pursuit methods

Peter J. Huber

Harvard University, Cambridge MA

ABSTRACT

Recently, Friedman and Tibshirani (1981) have proposed to use projection pursuit methods for multivariate density estimation. This paper develops a new method and technical issues in density estimation and describe several variations to projection pursuit density approximation and estimation. A new measure of fit between multivariate densities is introduced (maximum charged relative entropy), and some of its properties are derived. It seems to be particularly well adapted to projection pursuit density approximation.

DTIC FILE COPY
This work was facilitated in part by National Science Foundation Grant MCS-79-04685 and Office of Naval Research Contract N00014-80-C-0512.

Department of Statistics
Harvard University
Cambridge, MA 02138

PROJECTION PURSUIT METHOD
Approved for Release
Date 10-10-80

NOV 30 1981
OFFICE

81 17

DISCLAIMER NOTICE

**THIS DOCUMENT IS BEST QUALITY
PRACTICABLE. THE COPY FURNISHED
TO DTIC CONTAINED A SIGNIFICANT
NUMBER OF PAGES WHICH DO NOT
REPRODUCE LEGIBLY.**

1. Why density estimation?

Density estimates are used in two quite distinct contexts

- (1) to aid with the visual analysis of data
- (2) to serve as substitutes for the unknown true density.

One occasionally hears the claim that density estimates are not really needed, and that statisticians would do better by using estimated measures in an intelligent fashion than by devising arbitrary density estimates. How far is this claim justified?

With regard to (1), the criterion is clearly unambiguous: it is easier to perceive nonrandomity and to approximately locate the modes in a histogram with well-chosen bin width than in the estimated alternative. In a two-dimensional scatterplot, differences in the underlying density of about 30% tend to go unnoticed, but not so in a density estimate. Compare in particular Figs. 1 to 5 (pp. 45-49) of Fryer's discussion on the paper by Barrow, Kendall and Sadanow (1971).

With regard to (2), the criterion may contain some truth. For instance, in adaptive estimation of location it is certainly most transparent to describe the estimated version of the mean function $-f'/f$ as being the log-likelihood derivative of a smooth estimate \hat{f} of the density f , but the smoothing constant (kernel width) and other peculiarities of \hat{f} must be determined not by any of the density estimation criterion, but by standards derived from the ultimate adaptation requirement. Compare Jones (1972).

The underlying situation where the estimated measure (in the raw or in a smoothed form) substitutes for the unknown true distribution is the so-called bootstrap (cf. Efron 1977): we are given a single random sample (x_1, \dots, x_n) of size n from some unknown true distribution F , and we should like to learn about the sampling properties of a certain statistic $T_n = T_n(x_1, \dots, x_n)$. We do this by drawing N random samples of size m (with replacement) from the original measure F_n (or from a smoothed version thereof), and we base our conclusions on averages over the population consisting of the N values of T_m calculated from these random samples. Thus we shall want to keep the order of the Monte Carlo error $O(N^{-1/2})$

remained below the intrinsic error $O(n^{-1/2})$ caused by the finiteness of the original sample, we usually choose N larger than n (but not unreasonably so). Usually, we take $m=n$, since nearly we are interested in the properties of statistics calculated from all the available data, but also the limiting case $m \rightarrow \infty$ is of particular interest.

It happens that most statistics possess some built-in smoothing properties, evidenced by the fact that they tend to work reasonably well also with rounded and grouped data. As a consequence, it is readily adequate to use the raw, unsmoothed empirical measure F_n at least so long as m is small (say $m \leq n$). If m is substantially larger than n , we can get into serious trouble (an obvious example is furnished by the bootstrap estimate of the variance of the sample mean). More generally, the bootstrap performs poorly with the raw F_n whenever the statistic T_n is connected with local features of the distribution (e.g. for many statistics based on nearest neighbor techniques). Then it becomes essential to bootstrap from a wisely chosen density estimate. Compare in particular Silverman's recent (1981) paper on using bootstrap methods for estimating and testing the number of modes.

All approaches to density estimation involve at least one arbitrary smoothness parameter (the bin width in histograms, the kernel width in kernel smoothers, a lagrange multiplier in the parabolic likelihood approach, etc.). The optimal choice of this parameter depends on the unknown smoothness of the true underlying density, but also, and perhaps even more so on the objectives of the investigator. If he is using density estimates as an aid to visual data summary, it is important to by several choices of these parameters. Cross validation in the sense of W. Allen and W. Allen (1975) is a great tool, but -- like the automatic exposure control in modern cameras -- it should be used with circumspection, and must be overridden on occasion if we are looking for very rough spots in an otherwise very smooth curve, automatic cross-validation will tend to oversmooth and to smooth away the spots of interest.

2. One dimensional density estimates

In comparison to higher dimensional problems one dimensional density estimation is easy and straightforward. There are several good and reasonably well understood approaches (plan

histograms, kernel estimates, pseudoed likelihood estimates, plus some others, each with several variants (e.g. whether one chooses bins with constant width or with constant occupancy numbers in a histogram, and similarly for kernel estimates). See *Tapia and Thompson (1979)*.

The choice between these approaches often is more a matter of taste than a question of actual statistical performance. For our purposes, we shall need computationally fast estimates. The histogram clearly is fast, but it may not be smooth enough. Probably the fastest smooth estimate is obtained by first generating a histogram with too narrow bin width (e.g. equal to the resolution interval of the ultimate density estimate), and then smoothing this histogram by a single bin wide smoother. Compare Freedman and Diaconis (1981) for bin width recommendations.

3. Two and three-dimensional density estimates and their graphical representations

In two dimensions, kernel estimates still work well, and they lead to nice (and computer friendly efficient) graphical representations, namely to the ones familiar from cartography: smooth look-alike contours with coloring.

Two dimensional histograms, or more precisely, stereographic or perspective projections of kernel histograms, are widely used but have serious drawbacks: it is often difficult to make out which kernel value on a particular point (x, y) of the base plane, and small peaks can disappear behind a wall of high ones.

Pseudoed likelihood estimates and related approaches based on spline smoothing may have even more serious drawbacks for dimensions two and higher. In particular, they exhibit awkward minor kind phenomena (such as the support points). Moreover, they are expensive to compute.

Dimension 3 is the highest that allows direct visual representation of densities. Crystallographers for example routinely make 3 dimensional electron density maps by drawing contours on 2 dimensional sections -- on stacks of parallel picture sheets, more usually and more conveniently, the same kind of pictures are now generated on high power

graphic display devices. A rather possibility -- also dependent on kinematic display facilities -- is to use enhanced scatter plots: show large random samples from a density that has been modified in emphasize high density regions and to downplay low density regions. This coloring might for instance be achieved by replacing the original density f by a rescaled power f^a . This generates the "sharpening by erosion" of Tolley and Tolley (1980).

4. Multidimensional density estimates

In dimensions higher than 3, direct visualization of density estimates is impossible (or at least highly impractical), our visual cortex is too much geared to 3 dimensions. In addition, we run into the so-called "curse of dimensionality": high dimensional space is mostly empty. For instance, if points are uniformly distributed in the 10-dimensional unit ball, then a ball with radius $1/2$ only contains a fraction $2^{-10} \approx 1/1024$ of the points.

As a consequence, we must think hard about what we want to look at, and what our estimates are supposed to achieve when they are used as substitutes for the unknown true density.

Ordinarily, we would prefer projections to sections (but projections of "thick" sections might sometimes be interesting). It becomes debatable whether we should first identify the interesting projections (either by a priori considerations, or by projection pursuit methods applied to the pointcloud), and then use a low dimensional density estimate on the projected data, or whether we should first find a full p dimensional density estimate and then determine the interesting low dimensional marginals. With density estimates based on projection pursuit ideas, the two approaches are in fact very similar.

Density estimates may serve as powerful visual clustering methods. What kinds of features can we hope to detect in high dimensional data, and by what methods?

First, we need to aware that because of the "curse of dimensionality" even large regions of zero density are difficult or impossible to spot.

Small high density clumps on the other hand can be spotted by appropriate inward looking techniques (but not by standard kernel estimates with constant bin width). More generally,

ually, nearest neighbor techniques will be able to spot substructures consisting of points over-undersampled near low dimensional (not necessarily linear) submanifolds.

Proposed parent methods (to be discussed below, see also Hader 1981b) appear to be the first used so far only ones able to recognize substructures consisting of points undersampled near linear manifolds of low dimension (e.g. hyperplanes).

Linear substructures of intermediate dimensions (dimension near $p/2$ in a p -dimensional space, where p is large) may be even more difficult to spot. Δ_n (which) potentially works in finding a minimal spanning tree, breaking it into connected components by deleting all links crossing a certain height, and then doing related principal component analysis on the over periods (cf. Hader 1981a, Ch. II, and Usher and Li, 1981).

If we reject density estimates for visual inspection, it is usually better to use slight under-sampling, that if we plan to use them in a bootstrap, then it is usually better to stop on the "first", slightly oversampled sub, in particular if we want to check whether a particular function might have been a mean random effect. Compare Silverman (1980).

If a density estimate is to be used solely as a visual aid, it does not do much harm if it occasionally has negative values. But if it is to be inserted into a bootstrap or a similar procedure, we better make sure that it is a genuine probability density. Moreover, the representation of the density estimate then should be chosen such that it is easy to draw random samples from it.

I hope this short discussion has made it clear that we cannot expect any single approach to density estimation to do well under all circumstances.

2. Proposed parent density approximations

Proposed parent density estimates were recently proposed by Friedman and Stuetz (1981) as an extension of the projection pursuit regression idea. The basic idea is to search for a one dimensional projection for which the projection of the points and the marginal density of the original p dimensional density estimate disagree most, and then to improve the estimate

made by multiplying the marginal density in that particular direction. We shall now show that the Friedman-Stuetz approach solves a minimum problem posed in terms of relative entropy.

We first discuss a non-sampling version, namely the problem of approximating a density f in \mathbb{R}^p by a function g (if possible a probability density) having a specified functional form.

For example, we might approximate f by an additive decomposition of the form

$$f(x) \approx g^{(k)}(x) = \sum_{j=1}^k h_j(q_j^T x), \quad (5.1)$$

where $x \in \mathbb{R}^p$, $q_j \in \mathbb{R}^p$, and the h_j are functions of one real variable.

It is clear that under mild regularity conditions we can obtain arbitrarily accurate approximations of this form. Assume for instance that the characteristic function ϕ of f is absolutely integrable, then f has the harmonic representation

$$f(x) = (2\pi)^{-p} \int \hat{f}(s) e^{-is^T x} ds.$$

Any finite Riemann sum approximating the integral on the right hand side then yields an additive decomposition of the type (5.1).

Though the representation (5.1) has awkward features while $g^{(k)}$ converges to f uniformly on compacts when the mesh of the Riemann sum is extended and refined, the approximating sums clearly are not Lebesgue integrable, and they certainly do not exhibit the global features of a "typical" density (i.e. having a hump in the middle, and decaying to 0 as $|x| \rightarrow \infty$). We can restore integrability by multiplying the right hand side of (5.1) by a factor like the $\exp(-|x|^2)$, or equivalently, by picking a fixed, strictly positive, smooth probability density in \mathbb{R}^p (e.g. the standard Gaussian), and taking all densities relative to that instead of Lebesgue measure. Perhaps even worse, it is difficult to make g everywhere positive and to keep it positive and integrating to 1 while more terms are added to the sum.

Nevertheless, additive decompositions have technical advantages, the main one being that it is usually easy to calculate marginal densities.

The following is a somewhat more specific proposal for such a decomposition. First, we transform the distribution by an affine transformation such that it is centered at 0 and has unit

coordinates index (or alternatively, use a release procedure, such that the useful part of the distribution is retained at 0 and has unit variance). Then approximate f by a decomposition of the form

$$f(x) \approx g^{(k)}(x) = (2\pi)^{-p/2} e^{-\frac{1}{2} \sum_{j=1}^k \alpha_j^2 (x_j^2)} \quad (5.3)$$

with $\alpha_j = 1$. Note the formal similarity to an eigenvalue expansion.

Consequently, a multiplicative decomposition

$$f(x) \approx g^{(k)}(x) = \prod_{j=1}^k h_j(x_j^2) \quad (5.4)$$

looks more attractive. Note that if $k=p$, and if the α_j are linearly independent vectors then $g=g^{(p)}$ is a product density and can be written as

$$g(x) = \prod_{j=1}^p g_j(x_j) \quad (5.5)$$

in a suitable coordinate system, while the g_j are unidimensional probability densities. Clearly, such a decomposition is well suited to represent (approximate) product densities but it may yield awkward representations (i.e. involving too many terms) that are difficult to interpret, if the density f happens to be a mixture instead of an approximate product.

Now, we use an entropy decomposition $\log f$, and the density variance property of g . Another possibility is to write positively in to additively decompose \sqrt{f} . Incidentally, since the square root transform doubles the variance of the histogram, the latter approach also has some advantages from a sampling point of view.

Any scheme for iteratively determining directions α_j and successive approximations $g^{(k)}$ shall be called a *projection-based approximation*.

The quality of the approximation of g to f can be measured in many ways, e.g. by

(1) relative entropy

$$E(f, g) = \int \log(f/g) f \, dx,$$

or by

(2) Hellinger distance

$$H(f, g) = \int (\sqrt{f} - \sqrt{g})^2 dx,$$

or by any other measure of distance between distributions (e.g. Prokhorov distance, bounded Lipschitz metric, etc.). Since we are working with densities we are naturally more attracted to (1) or (2) than to the other distances mentioned. Among them, (1) is of course particularly well suited to an additive decomposition of $\log f$, (2) to an additive decomposition of \sqrt{f} .

Note that both Hellinger distance and relative entropy are invariant under affine transformations. Hellinger distance is a metric; relative entropy is not (it is not symmetric in its two arguments), but it follows from Jensen's inequality that $E(f, g) \leq 0$, and that $E(f, g) = 0$ implies that $f = g$ a.s.

We may allow $k < p$ in the decompositions. In the absence of judge factors g then fails to be Lebesgue integrable, but $E(f, g)$ still may be finite. However, in such cases it is usually preferable to pad g by $p-k$ "flat" (e.g. Gaussian) components in order to turn it into a genuine probability density.

6. Measurement of relative entropy

In this section we are concerned with optimal choices for the directions α_j and the factors h_j in the decomposition (5.4). Assume first that $k=p$ and that the g_j are fixed, linearly independent vectors in \mathbb{R}^p . We may choose g_j to be the j th coordinate direction without loss of generality. We first consider the problem of finding the best approximation in the relative entropy sense of the given density f by a product density

$$g(x) = \prod_{j=1}^p g_j(x_j),$$

where the g_j are unidimensional probability densities

(Relative entropy

$$E(f, g) = \int \log f - \sum_{j=1}^p \log g_j(x_j) f \, dx, \dots dx_j$$

is measured by increasing

$$= \sum_{j=1}^p \int_{a_j}^{b_j} f_j(x_j) f_1(x_1, \dots, x_p) dx_1 \dots dx_p,$$

which in turn is minimized by minimizing

$$= \int \log f_j(x_j) f_j(x_j) dx_j$$

for each j separately, which is

$$f_j(x_j) = \int f(x_1, \dots, x_p) dx_1 \dots dx_{j-1} dx_{j+1} \dots dx_p$$

is the j th marginal density.

Since $E(f_j, g_j) = 0$ for $g_j \perp f_j$, the minimum density is achieved for the unique choice

$$g_j = f_j$$

It is interesting to note that this calculation at the same time proves that Shannon entropy

$$E_X(f) = - \int \log(f) f dx$$

satisfies

$$E_X(f) = \sum_{j=1}^p E_X(f_j),$$

with equality if $f = \prod f_j$.

By letting the p dimensions x_j vary simultaneously and freely, we may more generally approximate f by the best possible product density (we do not worry about existence of the minimum for the minimum):

$$g(x) = \prod_{j=1}^p g_j(x_j)$$

It is not clear whether a discrete projection pursuit approach will achieve the best possible approximation by a product density; first solve a minimum problem to find g_j , then find the minimum for the minimum.

But if f is an exact product density in a suitable coordinate system, then a properly designed discrete approach will pick up the unique factors one at a time. This is a non-trivial result involving some subtle properties of entropy; it shall be sketched briefly.

It is convenient to use random variable terminology and to write $E_X(X) = E_X(f)$ if the one-dimensional random variable X has density f . Then the functional

$$Q(f) = E_X(-\log(f)) = \log \sigma(f) - E_X(f).$$

where $\sigma(f)$ is the standard deviation of X , is already invariant:

$$Q(X+Y) = Q(X) = Q(Y)$$

and has the property that for independent random variables X, Y :

$$Q(X+Y) \geq \max(Q(X), Q(Y)),$$

and that the inequality is strict unless both X and Y are normal, see Huber (1981b). Note that $Q(X) = E(f, g)$, where g is a normal density with the same mean and variance as f .

Now assume that f is a product density, without loss of generality we may assume that

$$f(x) = \prod f_j(x_j),$$

and that the relative entropies are in decreasing order:

$$E(f, g_1) \geq E(f, g_2) \geq \dots \geq E(f, g_p).$$

The above mentioned inequality for Q implies that the maximum of Q is reached at a factor of f , namely at the factor f_1 with the largest relative entropy $E(f, g_1)$. We now divide out this factor and replace f by

$$f^*(x) = f(x)g_1(x_1)/f_1(x_1)$$

Thus if f^* is subjected to the same process as f before, the second factor f_2 is picked out, and so on.

Playing around with a few synthetic examples has led to the following observations. Measurement of Shannon entropy $E_X(f)$ and of relative entropies seems to be a numerically unstable process if the covariance matrix of f is poorly conditioned (i.e. if the ratio between the largest and the smallest eigenvalue is large). This is made worse by the fact that there will be many bad minima in general, and that the relative entropies are deep but narrow even if the covariance matrix is well conditioned. So the minima may be difficult to find.

If f has finite second moments, then the best Gaussian approximation to f in the relative entropy sense has the same mean vector and covariance matrix as f . This is easily shown by a variational argument. In view of the preceding remarks it would therefore seem advisable to apply an affine transformation such that in the new coordinate system f has mean 0 and unit covariance matrix, and to start the approximation process with $g^{(0)} = \varphi$ being the standard p -

where f_n and g_n are the marginal densities of f and g respectively in direction n , and where n is chosen such that it minimizes

$$E(f, g) - E(f, g_n) = E(f_n, g_n) \quad (0.1)$$

(At the moment, we are not concerned about the existence of such an n ; minimization within a restricted relative error tolerance is in fact good enough for all practical purposes.)

The procedure just described shall be referred to as the *projection pursuit approximation method* in the following.

2. Convergence in maximum marginal relative entropy

We can use *maximum marginal relative entropy*

$$E^*(f, g) = \sup_n E(f_n, g_n)$$

as a measure of discrepancy between two p -dimensional densities f and g . Clearly, because of (0.1), $E^*(f, g) \leq E(f, g)$. Since any distribution is uniquely characterized by the set of its marginals in all possible directions (Cramér-Wold theorem), we have

$$E^*(f, g) = 0 \iff f = g \iff E(f, g) = 0.$$

Since each step of projection pursuit decreases $E(f, g)$ by $E^*(f, g)$, the latter quantity clearly converges to 0, and for any given $\epsilon > 0$, it takes at most $E(f, g)/\epsilon$ steps to reach a density \tilde{g} for which $E^*(f, \tilde{g}) \leq \epsilon$.

Maximum marginal relative entropy $E^*(f, g)$ seems to be a concept particularly well suited to projection pursuit approximation, and it deserves closer study. Let f be a fixed probability density, while g_n is an arbitrary sequence of probability densities

Clearly, $E(f, g_n) = 0$ implies $E^*(f, g_n) = 0$. I conjecture that the reverse implication is false, but I do not have a counterexample.

We now shall derive some consequences of E^* -convergence.

Proposition 2.1. $E^*(f, g_n) = 0$ implies that $g_n \rightarrow f$ in the sense of weak convergence of the underlying measures

dimensional normal distribution. This choice also advantageously provides the "best" peaking function h .

Let g be an approximation to the p -dimensional density f . We shall now attempt to improve this approximation by replacing $g(x)$ by $\tilde{g}(x) = g(x)h(x_1)$, where h depends on the first coordinate only. Note that f and g determine the same conditional density given x_1 . An intuitively attractive choice then is to determine h such that the marginal distribution of g in direction x_1 agrees with that of f . This indeed minimizes relative entropy.

Lemma 0.1. Relative entropy $E(f, g)$ is minimized by the choice

$$h(x_1) = f(x_1)/g(x_1),$$

where f_1 and g_1 are the marginal densities of f and g in direction x_1 .

Proof. We note that the conditional density, given x_1 , is the same for f and for g nearly

$$g(x_1) = g(x_2, \dots, x_p | x_1) = g(x)g(x_1),$$

and that $g_1(x_1)h(x_1)$ is the marginal density of \tilde{g} . Thus

$$\begin{aligned} E(\tilde{g}, g) &= \int (\log f - \log \tilde{g}) f \, dx \\ &= \int (\log f(x_1) - \log g(x_1) + \log f(x_1) - \log(g(x_1)h(x_1))) f \, dx, \end{aligned}$$

which is minimized by choosing

$$\int (\log f_1 - \log(g_1 h)) f \, dx = E(f, g h),$$

and this clearly is achieved for the unique choice $g h = f_1$.

For the minimizing choice of g we have

$$E(f, g) - E(f, \tilde{g}) = E(f, g h).$$

If we can choose the peaking function h , then the best possible improvement of relative entropy that can be achieved through replacing $g(x)$ by $\tilde{g}(x) = g(x)h(x_1)$ clearly is achieved with

$$h = f_1/g_1.$$

Immediately, this proposition is just another version of the Grönwall-Wold theorem, compare Hillebray (1981, p.49). We prove it with the aid of two auxiliary lemmas of independent interest.

Lemma 7.2. Let $z > -1$. Then

$$z - \log(1+z) \leq \frac{1}{4} \max(1, z^2).$$

Proof. For $z \leq 1$ the assertion is true, since it holds for $z=1$ and the left hand side is nondecreasing. We note that

$$t(z) = z - \log(1+z) = \frac{z^2}{4}$$

satisfies for $z \geq 0$ and its derivative

$$t'(z) = \frac{z(1+z)}{2(1+z)^2}$$

satisfies $t'(z) > 0$ for $-1 < z < 0$, $t'(z) > 0$ for $0 < z < 1$. Hence the assertion also holds for $z < -1$.

Lemma 7.3. For any two densities f and g with respect to Lebesgue measure in \mathbb{R}^p we have

$$\int |f-g| dx \leq 4 \|E(f, g)\|^{1/2}.$$

Proof. For $h=g-f \geq 1$, then

$$\begin{aligned} E(f, g) &= - \int \log(g/f) f dx \\ &= - \int \log(g/f) + g/f - 1 f dx + \int (1-g/f) f dx \\ &= \int (h - \log(1+h)) f dx + \int_{f=0} (f-g) dx \\ &\leq \frac{1}{4} \int \max(1, h^2) f dx \end{aligned}$$

by the Grönwall inequality (or more precisely, by repeating the proof) we obtain for $\delta > 1$

$$E_f(h) \leq \delta \int \frac{h^2(f, g)}{g^2}.$$

Assume now that $E(f, g) > 1/4$, and put $\delta = [4E(f, g)]^{1/2}$. Then

$$E_f(h) \leq \delta.$$

Let C be the set $\{x : |h(x)| \leq \delta\}$. Then $|g-f| \leq \delta f$ on C . We have

$$\int_C f dx \leq \delta.$$

and

$$\int_C g dx \leq \int_C (f-\delta f) dx \leq (1-\delta) \int_C f dx.$$

Hence

$$\int_C g dx \leq 2\delta.$$

Thus

$$\begin{aligned} \int |f-g| dx &= \int_C |f-g| dx + \int_{C^c} |f-g| dx \\ &\leq \delta \int_C f dx + \int_{C^c} f dx + \int_{C^c} g dx \leq 4\delta. \end{aligned}$$

This proves the lemma, since for $E(f, g) \leq 1/4$ the assertion is trivially true.

Corollary 7.4. Convergence in relative entropy $E(f, g_n) \rightarrow 0$ implies L_1 convergence (i.e. in total variation) and convergence in Hellinger distance.

Proof. L_1 convergence is an immediate consequence of Lemma 7.3, and convergence in Hellinger distance follows from the remark that $(\sqrt{f} - \sqrt{g})^2 \leq |f-g|$.

Proof of Proposition 7.1. In view of Lemma 7.3, $E^p(f, g_n) \rightarrow 0$ implies that the marginal densities converge uniformly in L_1 convergence:

$$\sup_x \int |U_n - g_{n,x}| dx \rightarrow 0.$$

Hence, the characteristic functions ψ_n of the marginals converge uniformly, and since the characteristic function ψ of any density f is related to the characteristic functions ψ_n of its marginals f_n by

$$\psi(t) = E[e^{itx}] = \psi_n(t),$$

it follows that the characteristic functions ψ_n of g_n converge uniformly to ψ .

$$|\psi(t) - \psi_n(t)| \leq \sup_x \int |U_n - g_{n,x}| dx$$

Hence, g_n converges weakly.

If f is sufficiently smooth, so that its characteristic function is absolutely integrable, and if

The sequence g_k is generated by the projection pursuit method described at the end of the preceding section. Then I conjecture that $g_k \rightarrow f$ uniformly and in the L_1 sense. Actually, I can prove only a very special case (which however covers the principal intended application):

Proposition 2. Assume that f is continuous with a compact support.

$$f = \int \phi \circ \varphi$$

where f is continuous, and φ is normal $N(0, \sigma^2)$. Then the sequence $g_k(k=0, 1, \dots)$ of approximating densities, constructed by projection pursuit as in Section 6, converges uniformly and in L_1 to f .

Proof. Each of the g_k allows a decomposition

$$g_k^{(1)} = g_k^{(2)} \circ \varphi$$

This is proved by induction. It clearly is true for $k=0$, since $g_0^{(1)}$ itself is Gaussian.

Now assume that $g = g^{(1)}$ can be decomposed. For the following argument it is essential that φ is a product density in all orthogonal coordinate systems, so that coordinates can be extended coordinatewise. Choose the coordinate system such that φ is the first coordinate direction. Then the decomposition

$$g = g(x_1)g(x_2)$$

shows the required action.

$g = g(x_1)g(x_2)$ is obtained by convolving their bivariate counterparts with $(p-1)$ -fold $g(x_1)$ and $g(x_2)$ taken respectively. If in the last displayed expression we add 1 dimensional normal distributions respectively. If we obtain the decomposition of the next term replace $g(x_1)$ by $f(x_1) = f(x_1)g(x_1)$, we obtain the decomposition of the next term $g^{(2)}(x)$.

The characteristic functions of $g^{(1)}$ and $g^{(2)}$ are related by $g^{(1)}(x) = g^{(2)}(x) \exp(-\sigma^2 |x|^2/2)$, and similarly for ϕ , hence ϕ and $g^{(1)}$ are separated by $\exp(-\sigma^2 |x|^2/2)$.

Thus the $\phi^{(k)}$ are absolutely independent, and $g^{(k)}$ can therefore be represented as

$$g^{(k)} = (2\pi)^{-p/2} \int \phi^{(k)}(x) e^{-\sigma^2 |x|^2/2} dx$$

Since the sequence $\phi^{(k)}$ converges uniformly, it now follows from the representation of ϕ and $g^{(k)}$ that the sequence

$$|f(x) - g^{(k)}(x)| \leq (2\pi)^{-p/2} \int |\phi(x) - \phi^{(k)}(x)| dx$$

converges uniformly to 0. L_1 convergence follows trivially.

6. Projection pursuit density estimation

It is straightforward to change the projection pursuit approximation procedure of Section 6 into a density estimator. The first step is to standardize the population in R^p by an affine transformation so that it is centered at 0 and that the covariance matrix is the unit matrix. Note that there are some very delicate robustness questions our density estimates should not be sensitive to occasional outliers, but they should be able to pick up long tails in the underlying distribution. These two requirements are in some sense contradictory, and we took a rational theory for dealing with them. But, from a pragmatic point of view, we note that all density estimates have trouble coping with isolated points in the tails (they tend to produce capely isolated bumps in the estimate), so it is probably wise to set aside those points, to disregard them in the estimation procedure, but to show them as remarkable points in the pictures we produce.

Since it seems to be better if the initial estimate has long heavy tails than if it has too high tails, we would seem to have the choice of either using the classical, non robust mean and covariance matrix together with a Gaussian $g^{(0)}$, or else robust location and covariance estimates together with a density $g^{(0)}$ that is heavier tailed than the Gaussian. I believe the simple first version to be good enough in most cases.

The search for density estimate $g^{(0)}$ thus ordinarily is the p -dimensional standard normal. The approximation steps now can be described as follows.

Let

$$g(x) = g^{(1)}(x) = g^{(0)}(x) h_1(u(x)) \quad h_1(u(x))$$

be the usual density estimate

According to Section 6, we now should determine a direction $\alpha = \alpha_{k+1}$ such that it maximizes $E(f_{\alpha} | g_k)$, and then put $h_{k+1} = f_{\alpha} g_k$.

For a given α , f_{α} is straightforward to calculate: project the given sample in direction α , yielding x_1, \dots, x_n , and then calculate a one-dimensional density estimate f_{α} based on $\{x_1, \dots, x_n\}$.

The projection g_k of the random density estimate g is a well defined, known quantity, and from the point of view of theory does not present any problems. However, we may run into trouble with its actual calculation, in particular since it has to be calculated inside a minimization loop.

Direct minimization algorithms almost certainly is too slow. There are several approximating Monte Carlo approaches. The first one is to replace g by a sample y_1, \dots, y_m from g , and then to estimate g_k in the same way as f_{α} . This variant may not be easy to implement (how does one sample efficiently from g ?). A second one is to take a random sample y_1, \dots, y_m from $g^{(0)}$ (i.e. from the p-dimensional standard normal distribution), to put $y_i = h_i(u_i^1, \dots, u_i^p)$, and to create a histogram with bin widths Δ and value

$$f_j = \sum_{i=1}^m I_{\Delta_j}(u_i^1, \dots, u_i^p) / m \Delta$$

for the j th bin. Then apply a kernel smoother to this histogram to obtain the estimate g_k .

This approach will be probably poor if $g-g^{(0)}$ has a narrow, but high density peak somewhere, because then the random sample from $g^{(0)}$ picks up few observations there that the value g_k is maximally estimated. If this happens, it may be best to choose temporarily the algorithm to locate the offending peak, to replace $g^{(0)}$ by a mixture of two differently scaled normals (one of them at the peak in question), and then to recast the projection pursuit algorithm with the new $g^{(0)}$.

We proposed here to use *marginally relative entropy* as the criterion to be maximized when determining a new projection direction. This certainly is the conceptually purest approach. But conceivably, other measures of dissimilarity might offer advantages from a sampling or

computational point of view, the matter needs further investigation.

2. Minimization of Hellinger distance:

Hellinger distance is more complicated to minimize than relative entropy. Assume that g is the current approximation to f , and that we want to find a function $h(x_1)$ of the first variable x_1 alone such that the Hellinger distance between the probability density $h \cdot f$ and

$$g(x) = g(x)h(x_1)$$

is minimized. Clearly, g and g determine the same conditional densities $g(x_2, \dots, x_p)$ say $q(x_2, \dots, x_p | x_1)$, so $h = g/g_1$ is the quotient of the marginal densities in direction x_1 . We can minimize $H(f, g)$ by maximizing

$$\int \sqrt{fg} \, dx = \int \int \sqrt{fg} \, dx_2 \dots dx_p \int \sqrt{g_1} \, dx_1$$

under the side condition

$$\int g_1 \, dx_1 = 1$$

We write for short

$$w(x_1) = \int \sqrt{fg} \, dx_2 \dots dx_p,$$

then the Schwarz inequality implies that

$$\int \sqrt{fg} \, dx = \int w \sqrt{g_1} \, dx_1 \leq \left\{ \int w^2 \, dx_1 \right\}^{1/2} \left\{ \int g_1 \, dx_1 \right\}^{1/2}.$$

with equality if $\sqrt{g_1}$ is proportional to w . Hence we should take

$$\sqrt{g_1} = c w,$$

with the constant c such that the side condition is satisfied, and the optimal h then is $h = g_1/g$.

We noted earlier that for Hellinger distance an additive decomposition of \sqrt{f} might be more genuine. Assume

$$\sqrt{f} = \sqrt{g} + \sum_{i=1}^k u_i(u_i^1(x_1))$$

In this case it is convenient to ignore the requirement that g should integrate to 1

If \sqrt{g} is the correct approximation, then the best approximation of the form

$$\sqrt{g} + u(x_1)$$

is obtained by minimizing

$$\int (\sqrt{f} - \sqrt{g} - u)^2 dx_1$$

which clearly is minimized by the choice

$$u(x_1) = \int (\sqrt{f} - \sqrt{g}) dx_2 = 0$$

In neither of the two problems the search for the optimal direction α of projection appears to allow a simple and intuitive formalization (as in the relative entropy case, where the relative entropy of the marginal densities had to be maximized).

In the sampling case, even greater inconveniences seem to arise, since it not at all clear whether the expectations for w and for u can be calculated in a computationally efficient fashion.

10. Consistency of projection percent density estimates

By reinterpreting some of the results obtained in Section 7, consistency of projection percent density estimates is almost trivial to prove. But the proof at the same time shows why consistency here is an abstract technical concept.

Assume that we are given a sample of size n in R^p , let

$$\mu = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$$

be the empirical measure. We now apply a spherically symmetric normal kernel smoother to obtain the density estimate

$$f = \mu * \sigma(\sigma^2)_{\sigma^2}$$

Note that the assigned density f_{α} of f in direction α is obtained by applying the one-dimensional kernel $N(0, \sigma^2)$ to the projection

$$\mu_{\alpha} = \frac{1}{n} \sum_{i=1}^n \delta_{x_i \cdot \alpha}$$

of the original data.

It follows that if we divide the projection percent estimate concerning relative entropy, using a kernel smoother with Gaussian kernel in the projection, it converges in the L^1 sense to the p -dimensional kernel estimate f .

Since f can be deconvoluted with a normal component, it follows from Proposition 7.5 that the projection percent estimate converges uniformly and in L^1 .

Since the p -dimensional kernel estimate f is consistent under very weak assumptions on the true underlying density f if σ tends to 0 slowly while the sample size n goes to ∞ , the projection percent estimate is consistent too, provided we iterate it enough so that it approximates f sufficiently closely.

This result is not very helpful, however. After all, the main reason for using the projection percent estimate is that the sample size is too small for a p -dimensional kernel estimate to make sense, and we certainly would not want to iterate the projection percent estimate so far that it approximates the latter. The following example may illustrate these issues.

Example. Take a sample of size n from the standard normal $N(0, I_p)$ in p dimensions. Note that an equal mixture of two normal densities $N(1, \sigma^2 I_p)$ is multimodal for all σ , besides other wise. Thus, we may say that two sample points are merged in the p -dimensional kernel estimate $f = \mu * N(0, \sigma^2 I_p)$, if their Euclidean distance is $\leq 2\sigma$, and that they are separated otherwise. Then the expected number of merged pairs can be calculated as

$$m = \frac{1}{2} n(n-1)q,$$

where

$$q = P(|x_1 - x_2| \leq 2\sigma) = \chi_p^2(2\sigma^2).$$

Numerically, we obtain with $p=10$, $n=10^6$, $\sigma=0.1$, that $m \approx 0.4$. In other words, in the p -dimensional kernel estimate all 10^6 points, except maybe 1 or 2 merged pairs, are still separated. The one-dimensional marginal estimates on the other hand will be quite smooth. Note that for one-dimensional estimates a kernel width like 10^{-6} minimizes the asymptotic mean square error, and that the constant k is such that for $n=10^6$, the choice $\sigma=0.1$ is very nearly optimal (cf

Wegman 1972, p. 584).

A somewhat more meaningful commentary result might be the following expected one:

Let f be the true underlying density and assume that it has been standardized such that its overall part is approximated by the standard normal $g^{(0)}$. Let

$$g^{(k)}(z) = g^{(0)}(z)h_1(u_1^k(z)) \cdots h_k(u_k^k(z))$$

be the k th population percent approximation to the true density f , obtained by minimizing mean squared relative entropy. Assume that the sequences u_1, u_2, \dots is uniquely determined and that $E^*(f, g^{(k)}) < \epsilon_k$ for some sequence ϵ_k . Let $u_k \rightarrow 0$ as the sequence obtained by projection percent density estimation from samples from f of size n , using a Gaussian $K(t, \sigma^2)$ kernel smoother on the one-dimensional projections with $\sigma_n^2 \rightarrow 0$ as the sample size $n \rightarrow \infty$ (the precise order of convergence needs to be specified). Then, for each fixed k , we would expect that $u_{k,n} \rightarrow u_k$ and $\lim_{n \rightarrow \infty} E^*(f, g^{(k)}) < \epsilon_k$, and that the convergence is much faster than the convergence of the p -dimensional kernel estimate.

Since we do not know the true f for u_k , except perhaps in synthetic sampling situations we would in practice need an analogue of Mallows's C_p statistic, telling us when to stop the projection percent approximation process.

The kernel width to be used for smoothing the one-dimensional marginals needs investigation. Presumably, one should use a width $h_n \sim 0$ approximately optimal with regard to expected mean square error for a three times differentiable density (even if the underlying density is expected to be somewhat rougher), and one should in general tend to oversmooth rather than to undersmooth.

11. Sampling questions

For most percent density estimation and relative entropy used as a distance measure these are practical sampling questions. This section discusses them in a preliminary fashion.

We would like to approximate the p -dimensional density f by a function g of a certain form. Since f is unknown, we need approximate an estimate \hat{f} instead. Then, in general, f is

not a genuine p -dimensional density estimate, but stands short for the collection of estimates \hat{f}_n of the one-dimensional projections f_n of f . By \bar{f} we denote the collection of expectations $\bar{f}_n = E(\hat{f}_n)$.

It is hardly realistic to aim for an "optimal" approximation. For reasons of computational expediency, the density estimates \hat{f}_n will have to be histogram estimates with constant bin width, smoothed with a simple smoother (possibly an undersmoothing one - that is, smoothing in the limit). But anyway, we should like to keep the total approximation error $g - f$ small, and then, common sense tells us that we should try to adjust the estimation and approximation procedures such that the bias $\bar{f} - f$, the random error $\hat{f} - \bar{f}$ and the approximation error $g - \bar{f}$ are all of about the same order of magnitude.

Let $E(\hat{f}_n, g_n)$ be relative entropy (or some other measure of discrepancy between one-dimensional densities) and put

$$E^*(f, g) = \sup_n E(\hat{f}_n, g_n),$$

$$E(f, g) = \inf_n E(\hat{f}_n, g_n).$$

Somewhat abstractly, we aim for the following approximate relations between bias, random error and approximation error. The bias should be smaller than the average random error, and the approximation error should be larger than the average random error, but it may fall below the mean random error. In formulas

$$E^*(f, \bar{f}) \approx E(\bar{f}, f) < E^*(g, f) \approx E^*(\hat{f}, f) \quad (11.1)$$

Before we discuss these errors further, we must state customary remarks of a more general nature. We note that if the difference $\Delta f = g - f$ is infinitesimally small, then a Taylor expansion gives for relative entropy

$$E(f, g) = E(f, f) + \frac{1}{2} \int \left(\frac{\Delta f}{f} \right)^2 f \, dx. \quad (11.2)$$

involving higher order terms. Thus, relative entropy is, essentially, a weighted average of the squared relative error. Compare also Lemma 7.2, and the proof of Lemma 7.3.

The above formula (11.2) indicates that spread error is needed in the tails of f , where f is small and the relative error may be unboundedly large. The problem can be partly alleviated by smoothing more heavily in the tails, but since this increases the relative bias, this does not help too much. Thus, we should truncate the tails, but make that the relative sampling error

$$E(\bar{f}, \bar{f}) \approx \frac{1}{2} \int \bar{f} \bar{f} \, d\mu$$

has the expectation

$$\frac{1}{2} \int \text{var}(f) \, d\mu, \quad (11.3)$$

and if we use a kernel estimate with kernel $K_\sigma(x) = \sigma^{-1} K(x/\sigma)$, that is

$$\bar{f}_n(x) = \frac{1}{n} \sum K_\sigma(x - x_i),$$

we obtain that asymptotically (for large n and small σ) the integrand of (11.3) is independent of x :

$$\text{var}(\bar{f}_n) \approx \frac{1}{n\sigma} \int K^2 d\mu.$$

Hence, truncating the range of integration is tricky: the computed value may be unstable and dependent on the choice of the truncation points.

It is an open question by how much of how links with the tails, plus perhaps a modification of $E(f, g)$ (see below) are able to mitigate these awkward features of relative entropy.

An alternative alternative to a kernel estimate with constant width is a histogram estimate with constant occupancy number n_j and variable bin width. Technically, it is somewhat easier to deal with histograms where the bin width has been adjusted such that the expected occupancy number $E(n_j) = n_j$ is the same for all k bins, with $q = 1/k$; asymptotically, both versions are equivalent. Then, the ratio between the estimated and the expected average density in bin j is $f_j - n_j/n_j$. This last quotient also makes sense in the end limit, where $\bar{f}, \bar{f} = 0, 0$ is unknown. For this estimate, the relative bias in the end limit is unboundedly large. But, at least,

the estimate has nice asymptotic sampling properties, as follows. For

$$q_j = (n_j/n)/\lambda(n_j),$$

then $\sum_j q_j = 0$, $E(q_j) = 0$, and $E(q_j^2) = (1-q)/\lambda(n_j)$.

For large n , the relative entropy bias is

$$\begin{aligned} E(\bar{f}, \bar{f}) &= \int \log \left| \frac{\bar{f}}{f} \right| \bar{f} \, d\mu \\ &= - \sum_j \log(1 + q_j) q_j \\ &\approx \frac{1}{2} \sum_j q_j^2 \approx E(f, f). \end{aligned}$$

For large n , this is stabilized like

$$\frac{k-1}{2k} \lambda k^{-1}$$

Hence, if $1 \ll k \ll n$, we have for one-dimensional distributions

$$E(\bar{f}, \bar{f}) \approx \frac{k-2}{2k} \lambda \frac{\sqrt{2}(k-1)}{2n}. \quad (11.4)$$

If this estimate is applied to the marginal distributions of a p -dimensional density, $E(\bar{f}_p, \bar{f}_p)$ is asymptotically distribution free; its distribution depends only on the sample size n and the number k of bins. Clearly, this distribution increases to about $E^*(\bar{f}, \bar{f})$, the asymptotic distribution of the latter quantity needs to be investigated.

A third approach, probably the most rational one, would be to minimize the expected (asymptotic) mean square error of each x , that is, to minimize the sum

$$E \left| \frac{f}{f} - \bar{f} \right|^2 + \left| \frac{\bar{f}}{f} - f \right|^2$$

of each x . Unfortunately, it is difficult to turn this into a computationally feasible estimate.

As already mentioned, it may be wise to modify the definition of $E(f, g)$ so as to make it less sensitive to tail effects. An alternative possibility is to replace the factor f/g by $(f+g)/(c+g)$, where $c > 0$ is a small constant depending on the sample size; in detail, we

If f itself is smooth, the approximation is valid directly for $u=0$, and we obtain that

$$f_u(x) \approx f(x) + \frac{u^2}{2} f''(x),$$

hence, according to (11.1),

$$E(f, f_u) \approx \frac{u^4}{8} \int \left| \frac{f''}{f} \right|^2 f \, dx.$$

The approximation error $E^*(g, f)$ is an observable quantity, so there is nothing to calculate; that we note that f is not estimating f itself, but a smoothed version thereof, so the following proposition about the relative entropies of smoothed densities is of some interest.

Proposition 11.2. Let f and g be probability densities on the real line, and let f_u and g_u be their respective convolutions with $N(0, u^2)$. Then, for $u > 0$,

$$\frac{d}{d(u^2)} E_{\Delta}(f_u) = \frac{1}{2} \int \left| \frac{f_u'}{f_u} \right|^2 f_u \, dx$$

and

$$\frac{d}{d(u^2)} E(f_u, g_u) = -\frac{1}{2} \int \left| \frac{f_u' - g_u'}{f_u - g_u} \right|^2 f_u \, dx,$$

(i.e. the derivatives are proportional to Fisher information and relative Fisher information respectively).

Proof. The proof is straightforward and uses Lemma 11.1 to differentiate the following expressions for entropy and relative entropy under the usual sign and differentiation rules: the result by parts

reduces to (11.4a) $E(f, g)$ by

$$E_g(f, g) = \int \left(\frac{f}{g} - 1 \right) \log \frac{f}{g} g \, dx; \quad (11.5)$$

note that $E_g = E$ if f dominates g , and that $E_g(f, g) > 0$ for $f \neq g$ (this follows easily from Lemma 7.2).

The problems just discussed are related to Shannon estimation: we are estimating a point in a high dimensional (infinite) space, and it pays to decrease the total error by leaving the estimator. A detailed theoretical investigation is still outstanding.

We now return to a discussion of the error terms in (11.1). We need crude, order-of-magnitude estimates; the random error term has already been discussed (we object modestly) in formulas (11.3) and (11.4). We only discuss the simplest case, namely kernel smoothers with a normal kernel; then the estimated \hat{f}_u are marginals of a Gaussian density f , and the marginal densities satisfy a simple partial differential equation, as follows.

Lemma 11.3. Let f be a probability density on the real line, and let f_u be the convolution of f with $N(0, u^2)$. Then f_u has bounded derivatives of all orders, and

$$\frac{d}{d(u^2)} f_u(x) = -\frac{1}{2} \frac{\partial^2}{\partial x^2} f_u(x), \quad u > 0$$

Proof. Essentially, the lemma states some well known properties of the heat equation. A single partial differential equation follows. We first note the the characteristic function of f_u is

$$\phi_u(t) = \exp(-\frac{1}{2} u^2 t^2) \phi(t),$$

where ϕ is the characteristic function of f ; hence $\phi^2 \phi_u$ is absolutely integrable for all t if $u > 0$, and thus all derivatives of f_u exist and are bounded. Let $f^2 \approx \sigma^2$, Δu^2 , and let Z be (mean) $N(0, \Delta u^2)$. Then

$$f_u(x) = E(f_u(x + Z)).$$

and if we take a Taylor expansion with remainder term inside the expectation, we obtain that

$$f_u(x) = f_u(x) + \frac{1}{2} f_u''(x) \Delta u^2 + o(\Delta u^2),$$

which proves the assertion.

References

- Pillay (1980), *Concepts of Probability Measures*, Wiley, New York.
- Levanon, D. Koppel and I. Shafar (1971), *Spline Transformations*, Thine, New Diagnostic.
- Ask for the Statistical Data Analysis (with discussion), *J. Royal Statist. Soc.*, B 32, 1-70.
- Z. Chen and G. Li (1981), Robust principal components and regression methods via projection pursuit, *Techn. Rep. Dept. of Statistics, Harvard University*.
- W. Jones (1976), *Robustness in Models*, another look at the problem, *Ann. Statist.*, 7, 1-28.
- D. Fienberg and P. Dawid (1980), On the Bayesian as a density estimator, *Techn. Rep. No. 100*, Dept. of Statistics, Stanford University.
- U. Fienberg and W. Stuetgen (1981), Projection pursuit methods for data analysis, *SIAC studies IEEE Trans. Computers* C-30, 881-888.
- P. J. Rousseeuw and J. W. Tibshirani (1974), A projection pursuit algorithm for exploratory data analysis, *IEEE Trans. Computers* C-23, 881-888.
- P. J. Rousseeuw (1981a), *Robust Statistics*, Wiley, New York.
- P. J. Rousseeuw (1981b), Projection pursuit, *Techn. Rep. Dept. of Statistics, Harvard University*.
- H. Devroye (1981), Using kernel density estimates to investigate multimodality, *J. Royal Statist. Soc. B*, 43, 305-318.
- C. J. G. (1976), Adaptive maximum likelihood estimators of a location parameter, *Ann. Statist.*, 4, 315-328.
- R. A. Fisher and J. R. Thompson (1970), *Nonparametric Probability Density Estimation*, John Wiley & Sons, New York.
- P. A. Tibshirani and J. W. Tibshirani (1980), Methods for direct and indirect graphical display for data sets and their derivatives, *Proc. of the Sheffield Conference*, ed. by V. Barnett, Wiley (to be published).
- G. W. Jones and S. W. Jones (1976), A completely adaptive kernel density estimation function by

cross-validation, *Comm. in Statistics*, 4, 1-17.

R. J. W. Jones (1972), Nonparametric probability density estimation I: A survey of available methods, *Technometrics*, 14, 333-347.

